基于形态的时间序列相似性度量研究

董晓莉　顾成奎　王正欧
(天津大学系统工程研究所　天津　300072)

摘　要：时间序列重新描述和相似性度量是时间序列数据挖掘的研究基础，对提高挖掘任务的效率和准确性至关重要。该文提出了一种新的基于形态的时间序列符号描述，并给出相应的距离公式，以度量时间序列的相似性。该方法直观简洁，对数据的平移、伸缩不敏感，能够反映序列趋势变化的程度、去除噪声的影响，满足时间多分辨率要求。仿真结果表明，该方法具有较好的聚类性能，可以在不同分辨率下有效度量时间序列的形态相似性。

关键词：时间序列；数据挖掘；相似性度量；重新描述
中图分类号：TP311　　**文献标识码**：A　　**文章编号**：1009-5896(2007)05-1228-04

Research on Shape-Based Time Series Similarity Measure

Dong Xiao-li　Gu Cheng-kui　Wang Zheng-ou
(*Institute of Systems Engineering*, *Tianjin University*, *Tianjin* 300072, *China*)

Abstract: The representation and similarity measure of time series are the basis of time series research, which is quite important to improving the efficiency and accuracy of the time series data mining. This paper proposes a shape-based discrete symbolic representation and its corresponding distance measure to measure the similarity between time series. The present method is intuitive and compact, and not sensitive to the shifting, amplitude scaling, compression and stretch of data. The method can reflect the degree of the dynamic change of the tendency and erase the influence of the noises, and it has multi-scale characterization. The experimental results show that the approach has good effect in clustering，which can measure the shape-similarity of time series effectively under various analyzing frequency.

Key words: Time series; Data mining; Similarity measure; Representation

1　引言

时间序列广泛存在于经济活动和科学应用中。相似性研究是时间序列数据挖掘的一个最基本而比较困难的问题。正确、简洁的离散化原有序列，并正确度量其相似性是提高挖掘效率和效果的关键，也是进一步处理如：关联分析、分类与预测、聚类分析、异类分析及演化分析的的基础[1]。文献[2]提出把序列分成有意义的子序列，并使用他们的实际值函数进行表示。文献[3，4]提出把连续的时间序列离散为具有特定含义的符号化表示。但目前使用的这些相似性度量的方法大都是基于欧氏距离的，存在下列缺陷：(1)不具有形态识别能力；(2)无法有效体现动态变化趋势的相似性，如图 1 所示，A与B的形态变化相反，与C的形态变化相同，但基于欧氏距离的计算，会认为A，B的相似性大于A，C；(3)不能识别时间序列在不同分辨率下的模式变化[2]。文献[5]提出了模式距离度量相似性，但该方法只能反映动态趋势的绝对变化，对数据的纵向压缩、拉伸敏感，模式分类粗糙，不能区分图 2 所示的两类曲线。

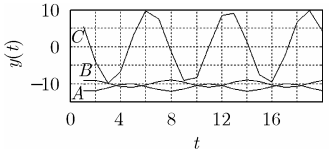

图 1　欧氏距离存在的问题图示

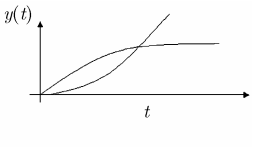

图 2　文献[3]不能区分的两条曲线

本文在时间序列分段线性化表示的基础上，提出“形态变化度量”的概念，并定义了“形态距离”公式，符合人们的视觉直观判断。本方法对数据的平移、伸缩不敏感；能够反映趋势动态变化的程度；模式分类细致，有利于提高检索、预测及分类、聚类精度；满足时间多分辨率要求，能够去除噪声的影响。

2　基于形态的时间序列重新描述

2.1　分段线性化和确定模式区分阈值

长度为L的时间序列的n段PLR模型表示为S，如式(1)所示[6]，y_{iL}, y_{iR} ($i=1,2,\cdots,n$)分别表示第i段的起始值和终止值，表示第i段结束的时间，n表示整个时间序列划分的直线段数目，$t_n=L$。

$$S=\{(y_{1L},y_{1R},t_1),(y_{2L},y_{2R},t_2),\cdots,(y_{iL},y_{iR},t_i),\cdots,(y_{nL},y_{nR},t_n)\} \tag{1}$$

2005-10-17 收到，2006-07-20 改回
国家自然科学基金(60275020)资助课题

模式区分阈值 th 可根据需要主观确定，可以是均值或中值的 10%左右，一般取 0.05~0.2 之间。

2.2 标准化处理

在保持形态不变的同时，经过标准化处理，使分段后的数据落在[0,1]之间，即 $nd=(y-\min y)/(\max y-\min y)$，$\mathrm{th}\to \mathrm{th}/(\max y-\min y)$。

2.3 形态描述

首先判断第 1 段的斜率，然后依次逐段比较斜率，确定各段的模式，如表 1 所示，其中$\Delta k=k_{(i+1)}-k_i$。本文将模式的变化表示为七元集合{快速下降，保持下降、平缓下降，水平，平缓上升，保持上升，快速上升}，将上述模式对应表示为$\boldsymbol{M}$={−3，−2，−1，0，1，2，3}，见图 3。使用者可以根据需要进一步细分或泛化，遵循“模式差异大，则数字距离大”的原则，使用不同的数字表示。$m\in\boldsymbol{M}$表示模式集合中的一个元素，一个时间序列S的形态可以表示为(模式，时刻)对的形式，见式(2)。其中：$m_i\in M$，$i=1,2,\cdots,n$；$t_i\cdots t_n$ 为该段的结束时间；n 为时间序列的分段数。算法伪代码见图 4，T_i表示S中的第i个分段(y_{iL},y_{iR},t_i)，$K(T_i)$表示S中第i个分段的斜率。

$$\breve{S}=\{(m_1,t_1),\cdots,(m_2,t_2),\cdots,(m_n,t_n)\} \tag{2}$$

表 1　形态模式列表

	$k_{(i+1)}<-\mathrm{th}$			$-\mathrm{th}<k_{(i+1)}<\mathrm{th}$	$k_{(i+1)}>\mathrm{th}$		
$k_i<-\mathrm{th}$	$\Delta k<0$	$\Delta k=0$	$\Delta k>0$	0	3		
	−3	−2	−1				
$-\mathrm{th}<k_i<\mathrm{th}$	−3			0	3		
$k_i>\mathrm{th}$	−3			0	$\Delta k<0$	$\Delta k=0$	$\Delta k>0$
					1	2	3

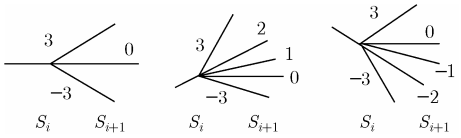

图 3　七元模式图示

2.4 齐序列处理

齐序列是指两个序列中对应的每一个模式的开始及结束时间相等，可以采用相互投影法得到[5]。本文对文献[5]的算法进行了改进，可同时将多个时间序列变为齐序列，效率更高。算法伪代码如图 5 所示，$\breve{\boldsymbol{S}}$是基于形态表示的时间序列$\breve{S}_1,\breve{S}_2,\cdots,\breve{S}_n$形成的数组，$\breve{\boldsymbol{S}}=[\breve{S}_1,\breve{S}_2,\cdots,\breve{S}_n]$，$T$是$\breve{\boldsymbol{S}}$中每个分段结束时间形成的数组，$\boldsymbol{M}$是处理后的齐序列数组。

3　时间序列的形态距离

时间序列的总长度是L，共有齐序列n段，每一段的作用时间为t_{ih}，在这一时间段内作用强度的变化为A_{ih}，即$t_{ih}=t(i+1)-t(i)$，$L=\sum_{i=1}^{n}t_{ih}$，$A_{ih}=y_{ir}-y_{il}$，M_i是经过

输入：(1)PLR 表示的时间序列 S; (2)模式区分阈值 th(th>0)。
输出：基于形态表示的时间序列 $\breve{S}$，如式(2)所示。

```
Š=F
S= (S − min(S)) / (max(S) − min(S))   //标准化处理
th= th /[(max(S)−min(S)]
calculate_K(T_i)    //计算各段的模式值
if K(T_1)<−th, Š(1)=3
    elseif K(T_1)>th, Š(1)=−3
    else Š(1)=0
end
for i=2:n
    if {K(T_{i−1})<−th or −th <K(T_{i−1})<th} & K(T_i)< K(T_{i−1}), Š(i)= −3; end
    if K(T_{i−1})<−th & K(T_i)= K(T_{i−1}), Š(i)=−2; end
    if K(T_{i−1})<−th & K(T_{i−1}) <K(T_i)< −th, Š(i)=−1; end
    if K(T_{i−1})>th & 0<K(T_i)< K(T_{i−1}), Š(i)=1; end
    if K(T_{i−1})>th & K(T_i)= K(T_{i−1}), Š(i)=2; end
    if {K(T_{i−1})>th or −th <K(T_{i−1})<th }& K(T_i)> K(T_{i−1}),
    Š(i)=3
    else Š(i)=0; end
end
```

图 4　形态描述算法伪代码

```
输入：基于形态表示的时间序列Š_1, Š_2,⋯, Š_n
输出：齐序列M_1, M_2,⋯, M_n
value=[ ]; t=0; i=1;
while t < max(T)
  value(i) =min(T)
  i=i+1; t=value（i）
end
for k=1:size(T,1)
    while n<=size(T,2)
        if value(n)=T(k,m), M(k,n)=mk(k,m);
        elseif value(n)< T(k,m),M(k,n)=mk(k,(m−1));
        else M(k,n)=mk(k,m);
        end
    end
end
```

图 5　齐序列对齐算法伪代码

齐序列处理后序列形态的符号表示。两个时间序列之间的形态距离公式为

$$D(M_1,M_2)=\frac{1}{L}\sum_{i=1}^{n}t_{ih}\times\left|A_{1ih}-A_{2ih}\right|\times\left|M_{1i}-M_{2i}\right| \tag{3}$$

两个时间序列的形态距离越小，两个时间序列之间的形态越接近。显然，该距离公式具有下列特点：

(1)保持时间加权特性。t_{ih}是第i个趋势的保持时间，对不同形态的保持时间进行加权。

(2)形态描述相同的时间序列之间的形态距离，与作用强度无关。

(3)形态描述不同的时间序列，作用强度差异越大，形态距离越大。

4 仿真

本文采用人工数据和股票数据进行仿真，使用经典的层次聚类法[7]，与欧氏距离、文献[5]的模式距离进行比较，从聚类时间和聚类效果两方面来验证本方法的有效性。

4.1 人工合成数据

设人工合成时间序列由函数$y(t)$来表示，则基本变形可定义为下列一组函数：(1)噪声NO($y(t)$= $y(t)$+$e(t)$)，$e(t)$为随机噪声。(2)平移SH($y(t)$)=$y(t)$+c_s，c_s为常数。(3)幅值按比例缩放AS($y(t)$)=$c_A y(t)$，c_A为正常数。(4)时间轴按比例缩放TS($y(t)$)=$y(c_1 t)$，c_1为正常数。(5)线性移动LD($y(t)$)=$y(t)$+$L(t)$，$L(t)$为直线方程。这些基本函数可以进一步组合出多个变形。

取$y_1(t)$=sin(t)，和$y_2(t)$=cos(t)，通过在 0~2π之间平均产生 200 个数，显示上述变形情况。各种变形的参数如表 2 所示，每次变化一个变形参数的取值。针对上述数据进行聚类，采用文献[8]定义的平均准确率法，平均准确率AA=(PA+NA)/2，积极准确率PA=Na/(Na+Nc)，消极准确率NA=N/(Nd+Nb)，其中，Na，Nb，Nc，Nd的取值分别是符合表 3 中a，b，c，d关系的时间序列的数量，仿真结果如表 4 所示。本方法能够对前四种变形正确聚类，当线性移动中的直线方程斜率较小时也能正确聚类。欧氏距离聚类速度较快，但精度较差，本方法聚类速度与模式距离相当，但聚类精度有显著提高。

表 2 合成时间序列的变形参数

变形参数
(1)噪声 附加上零均值的高斯白噪声 标准差是 0.01，0.10，0.15
(2)平移 平移常熟 c_s取 0.2，0.5，1
(3)幅值按比例缩放 缩放常数 c_A为 0.5，1.5，2
(4)时间轴按比例缩放 缩放常数 c_t为 0.5，0.8，1.5
(5)线性移动 直线方程为$y_L(t)$=y_0+$k_L(t)$， 其中y_0=0.2，k_L=0.2，0.5，1

表 3 时间序列的 3 种关系

算法聚类中属于同一类	手工分类中属于同一类	标识
是	是	a
是	否	b
否	是	c
否	否	d

表 4 基于 3 种距离公式的聚类方法比较

	欧氏距离	模式距离	形态距离
平均运行时间(s)	0.297	0.38	0.378
聚类精度(%)	64.3%	71%	90%

4.2 股票时间序列

采用文献[5]中的仿真实验数据，从 1986 年 6 月 6 日开始的 2700 个工作日的 6 种股票指数：SNGALLS，JAPDOWA，HNGKNGI，FTSE100，DAXINDX，AMSTEOE[9]。分别对 6 支股票的短期趋势(约 3 天)、中期趋势(约 15 天)和长期趋势(约 54 天)进行比较，聚类结果如图 6~图 12 所示。从聚类结果可以看出：

(1)本方法和模式距离均具有多分辨率特性，而欧氏距离不具有多分辨率特性，不能刻画不同时间尺度下时间序列的趋势相似程度。

(2)本方法较符合直观判断，而欧氏距离和模式距离的聚类结果直观性较差。如曲线 2 与其他 5 条曲线的形态差异较大，在本方法的所有聚类结果中，曲线 2 都被聚在最外层，而其他两种方法的聚类结果中，却将其与其他 3 条曲线聚在一起。

显然，本文方法能够有效体现时间序列的动态特性，可识别时间序列在不同分辨率下的模式变化，且聚类准确率较高。

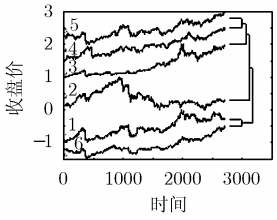

图 6 基于欧氏距离的聚类结果

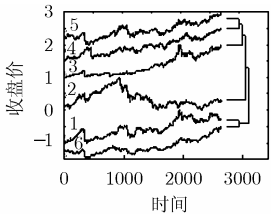

图 7 基于模式距离的聚类结果(900 段)

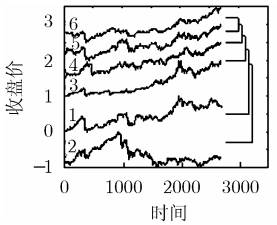

图 8 基于模式距离的聚类结果(180 段)

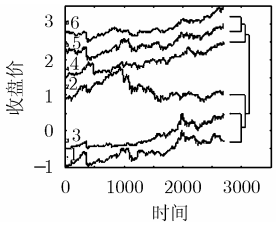

图 9 基于模式距离的聚类结果(50 段)

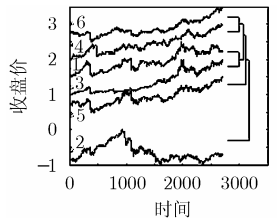

图 10 基于形态距离的聚类结果(900 段)

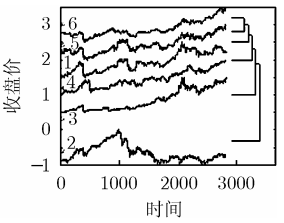

图 11 基于形态距离的聚类结果(180 段)

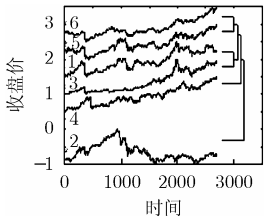

图 12　基于形态距离的聚类结果(50 段)

5　结束语

时间序列重新描述和相似性度量是时间序列挖掘任务的研究基础，本文提出了一种新的基于形态的时间序列相似性度量方法。本方法可定量描述时间序列的形态变化，直观简洁，且可避免标准化带来的形态变形，保证了标准化前后形态描述的一致性。人工数据和实际数据的聚类实验表明，本方法有较高的聚类精度和多分辨率特性，具有较好的应用前景。

参 考 文 献

[1] Chung Fu-Lai, Fu Tak-Chung, Ng V, and Luk Rt W P. An evolutionary approach to pattern-based time series segmentation. *IEEE Trans. on Evolutionary Computation*, 2004, 8(5): 471–489.

[2] Shatkay H and Zkonik S B. Approximate queries and representations for large data sequences. in Proc. Int. Conf. Data Engeering. Los Alamitos, CA: IEEE Computer Society Press, 1996: 536–545.

[3] Das G, Lin K I, and Mannila H. Rule discovery from time series, in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, New York City, 1998: 16–22.

[4] Kai O Y, Jia W, Zhou P, and Meng X. A new approach to transforming time series into symbolic sequences, in Proc.1st Joint BMES/EMBS Conf., Atlanta, GA, USA Oct. 1999, vol.2: 974.

[5] 王达, 荣刚. 时间序列的模式距离. 浙江大学学报(工学版), 2004, 38(7): 795–798.

[6] Keogh E and Pazzani M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In proceedings of the 4th Int'l Conference on Knowledge Discovery and Data Mining. New York, NY, Aug 27–31. 1998: 239–241.

[7] 方开泰, 潘恩沛. 聚类分析, 北京: 地质出版社, 1982: 44–51.

[8] 姜宁, 史忠植. 文本聚类中的贝叶斯后验模型选择方法[J]. 计算机研究与发展, 2002, 5: 580–587.
Jiang Ning, Shi Zhong-zhi. Bayesian posteriori model selection for text clustering. *Journal of Computer Research and Development*, 2002, 5: 580–587.

[9] Hyndman R J. http://www-personal.buseco.monash.edu.au/~hyndman/tsdl/data/FVD1.dat.2002–12.

董晓莉：　女，1975 年生，博士生，研究方向为数据挖掘、时间序列.

顾成奎：　男，1974 年生，博士，主要研究方向为系统建模与分析、数据挖掘、时间序列.

王正欧：　男，1938 年生，博士生导师，主要研究方向为数据挖掘、神经网络、系统辨识、人工智能等.